

Explainable AI: Moving from Numbers to Meaningful Insights via Prototype-Based Explanations

Jacek Karolczak

Poznan University of Technology
Graylight Imaging

Confession

Disclaimer: I wish this story did not happen. Unfortunately it did.



Confession

Disclaimer: I wish this story did not happen. Unfortunately it did.



One week, during a team meeting, I presented feature importance scores for a model I had trained.

My coworker looked at them and said:

These scores don't make sense. We need a better model.

Confession

Disclaimer: I wish this story did not happen. Unfortunately it did.



One week, during a team meeting, I presented feature importance scores for a model I had trained.

My coworker looked at them and said:

These scores don't make sense. We need a better model.



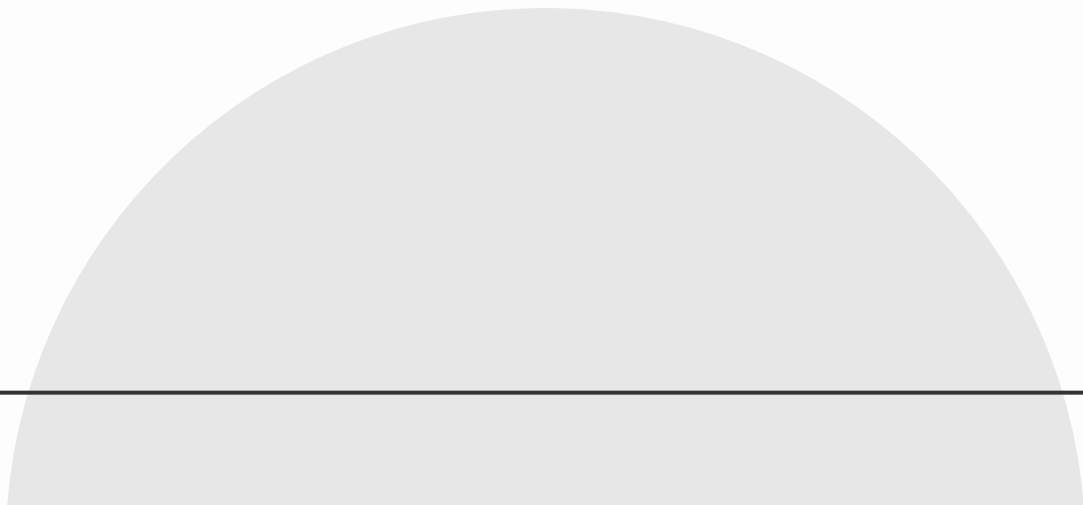
The following week, I showed feature importance scores for the same model - but this time, estimated using a different method.

He smiled and said:

This looks perfect. Let's move forward with this model.

At that moment, I asked myself:

**Maybe feature importance methods
are not actually that useful after all?**



saliency masks on the level of pixels are often unsuited for laypersons

- [9] Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Ser, J. D., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Paez, A., Samek, W., Schneider, J., Speith, T. and Stumpf, S. 2024. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *An International Journal on Information Fusion*, 106(102301), 102301. doi:10.1016/j.inffus.2024.102301
- [12] Speith, T. 2022. How to evaluate explainability – a case for three criteria. Proceedings of the 30th IEEE International Requirements Engineering Conference Workshops, REW 2022. pp. 92-97. doi:10.1109/REW56159.2022.00024

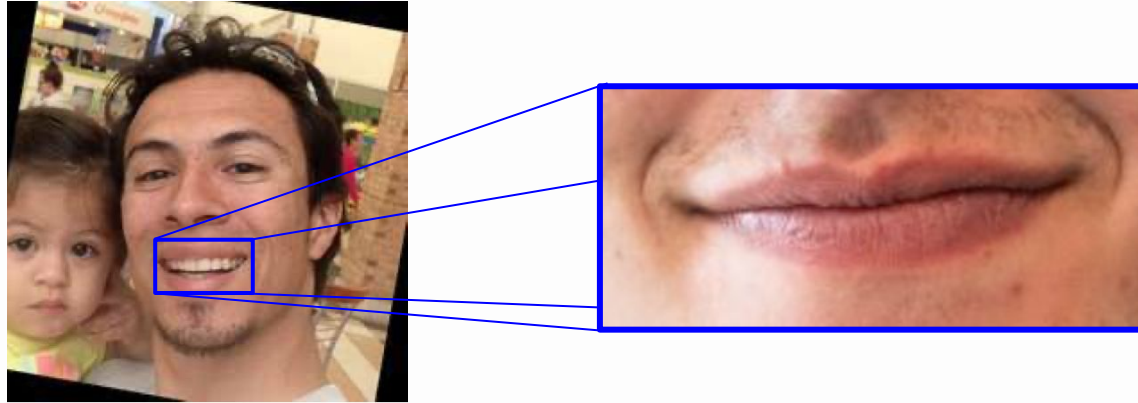
Issues with pixel attribution



Model classification: Young Adult

Saliency map says: It is the teeth hue!

Issues with pixel attribution



Model classification: Young Adult



~~Saliency map says:~~ It is the teeth hue!

Reality: The smile gave it away.

So what?

Find meaningful concepts. Do not get distracted by pretty heatmaps or isolated numbers.

Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions

Luca Longo^{1,2}  , Mario Brcic³, Federico Cabitza^{4,5}, Jaesik Choi^{6,7}, Roberto Confalonieri⁸, Javier Del Ser^{9,10,11}, Riccardo Guidotti¹², Yoichi Hayashi¹³, Francisco Herrera¹¹, Andreas Holzinger¹⁴, Richard Jiang¹⁵, Hassan Khosravi¹⁶, Freddy Lecue¹⁷, Gianclaudio Malgieri¹⁸, Andrés Páez^{19,20}, Wojciech Samek^{21,22,23}, Johannes Schneider²⁴, Timo Speith^{25,26}, Simone Stumpf²⁷

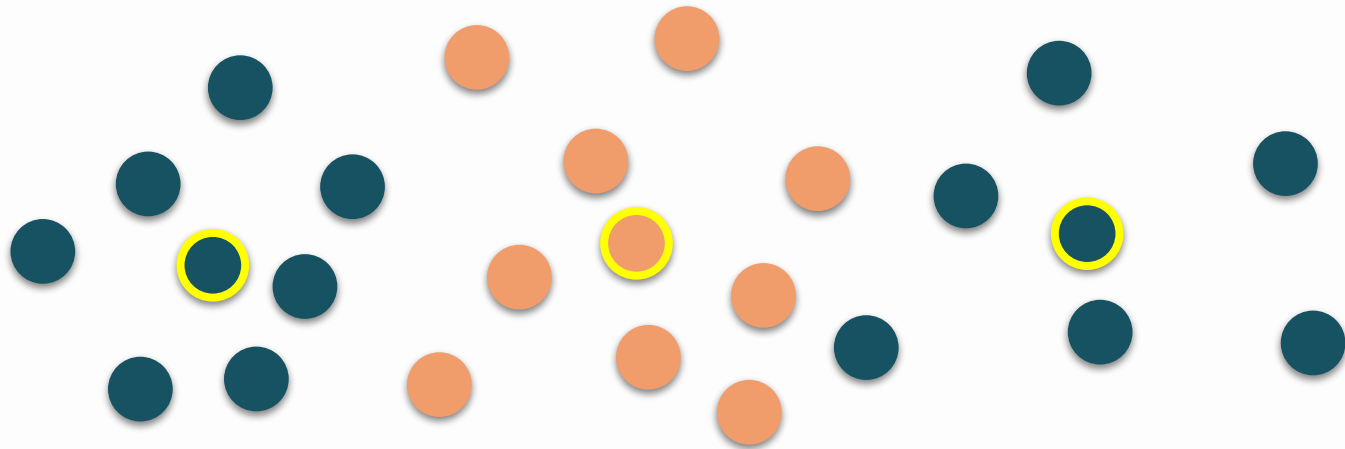
Abstract




Understanding black box models has become paramount as systems based on opaque Artificial Intelligence (AI) continue to flourish in diverse real-world applications. In response, Explainable AI (XAI) has emerged as a field of research with practical and ethical benefits across various domains. This paper highlights the advancements in XAI and its application in real-world scenarios and addresses the ongoing challenges within XAI, emphasizing the need for broader perspectives and collaborative efforts. We bring together experts from diverse fields to identify open problems, striving to synchronize research agendas and accelerate XAI in practical applications. By fostering collaborative discussion and interdisciplinary cooperation, we aim to propel XAI forward, contributing to its continued success. We aim to develop a comprehensive proposal for advancing XAI. To achieve this goal, we present a manifesto of 28 open problems categorized into nine categories. These challenges encapsulate the complexities and nuances of XAI and offer a road map for future research. For each problem, we provide promising research directions in the hope of harnessing the collective intelligence of interested stakeholders.

Audiences without technical background are often concerned with concepts, not with data

[9] Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Ser, J. D., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Paez, A., Samek, W., Schneider, J., Speith, T. and Stumpf, S. 2024. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *An International Journal on Information Fusion*, 106(102301), 102301. doi:10.1016/j.inffus.2024.102301

Global prototype explanation



The instances    summarize class distributions – they help us grasp what each class looks like to the model.

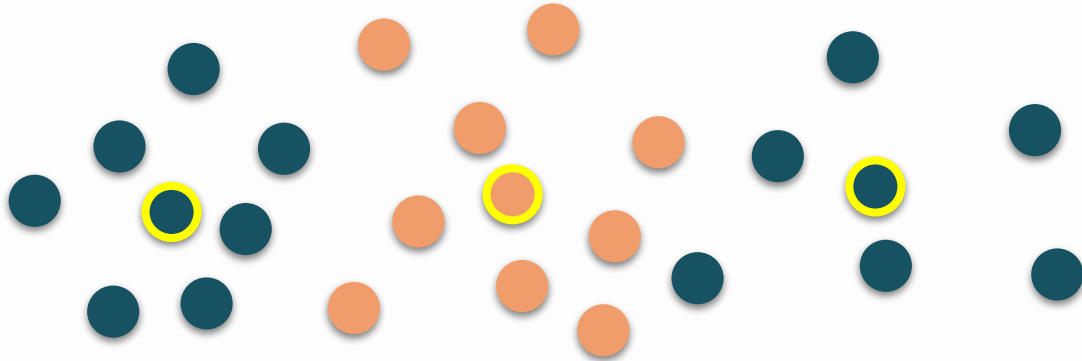
Local prototype explanation



The instance  belongs to class  because
is more similar to  than to 

Finding prototypes \approx solving a k-Medoid problem

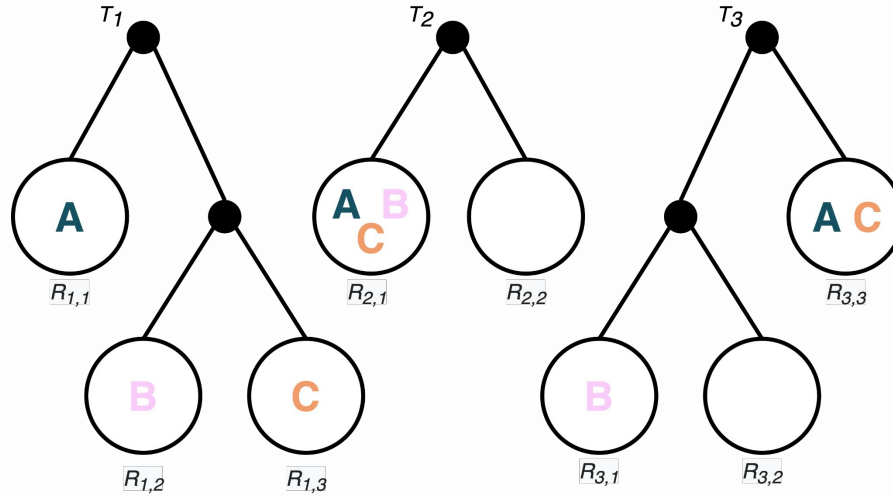
Finding prototypes is similar to solving a k-medoid problem – In both cases, one wants to select real instances that best represent the data by minimizing distances to other points.



How to quantify distance between two instances?

By measuring distances in embedding spaces or other internal model-specific representations.

Tree-space distance



$$d^{\text{TE}}(\text{A}, \text{B}) = 1 - \frac{1}{3} = \frac{2}{3}$$

$$d^{\text{TE}}(\text{B}, \text{C}) = 1 - \frac{1}{3} = \frac{2}{3}$$

$$d^{\text{TE}}(\text{A}, \text{C}) = 1 - \frac{2}{3} = \frac{1}{3}$$

[2] Breiman, L., 2001. Random forests. Machine Learning 45, 5-32. doi:10.1023/A:1010933404324.

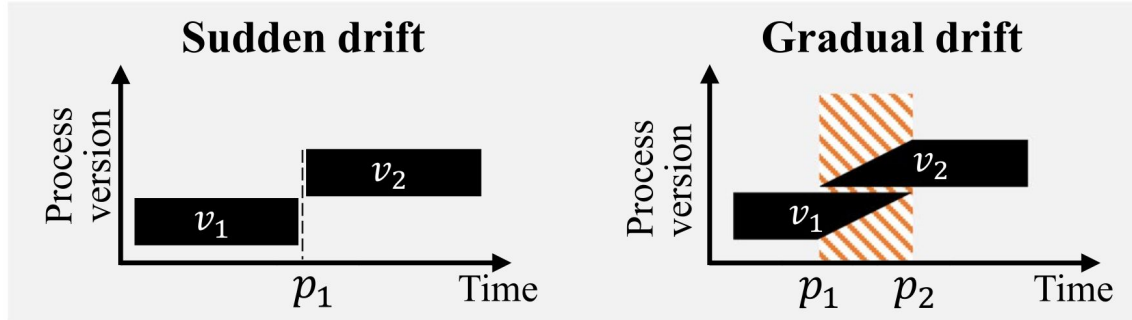
[5] Karolczak, J. and Stefanowski, J. 2024. A-PETE: Adaptive Prototype Explanations of Tree Ensembles. Progress in Polish Artificial Intelligence Research 5 : Proceedings of the 5th Polish Conference on Artificial Intelligence (PP-RAI'2024), 2-8. doi:10.17388/WUT.2024.0002.MiNI

Sounds easy, right?

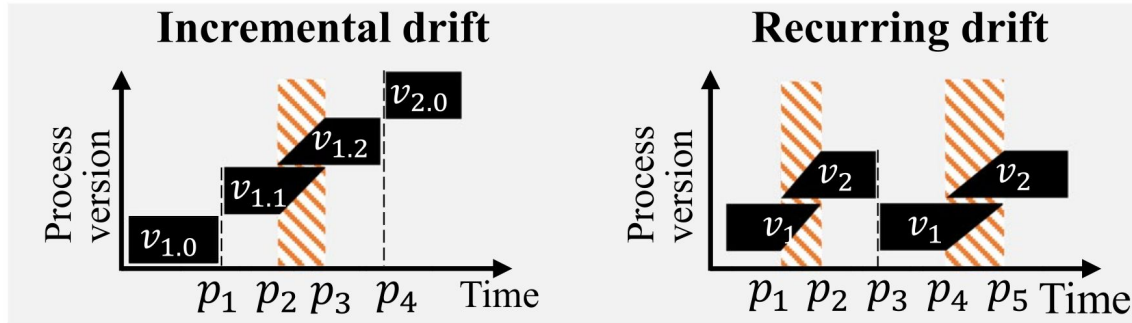
**The real world is messier
than dots on a whiteboard**

Data may change over time

Simple drifts

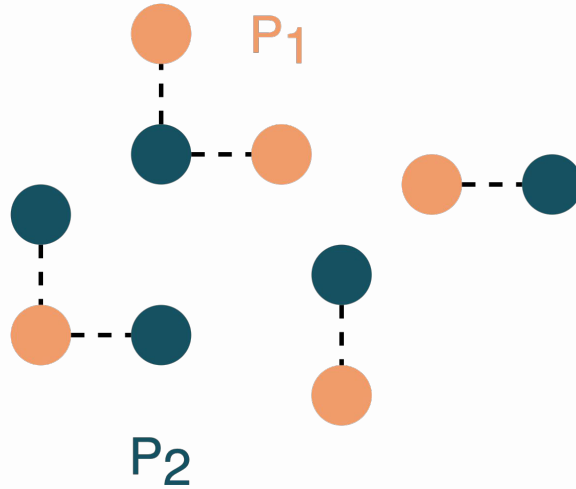


Complex drifts

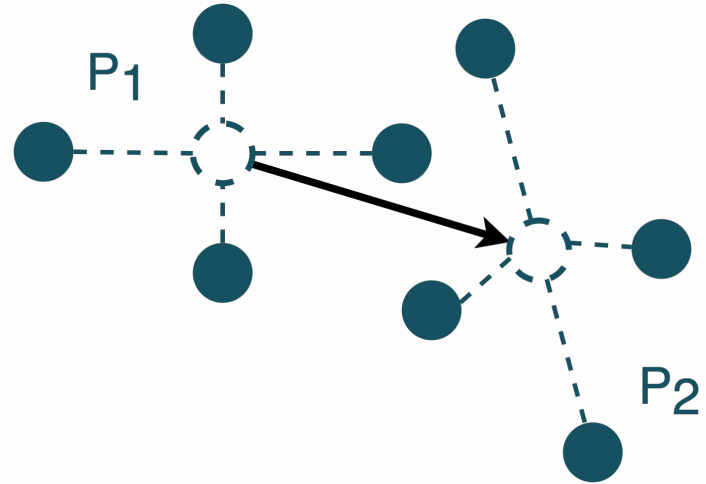


| - Sudden drift ▨ - Gradual drift

Measuring difference between two sets of prototypes



Mean minimal distance



Mean centroid displacement

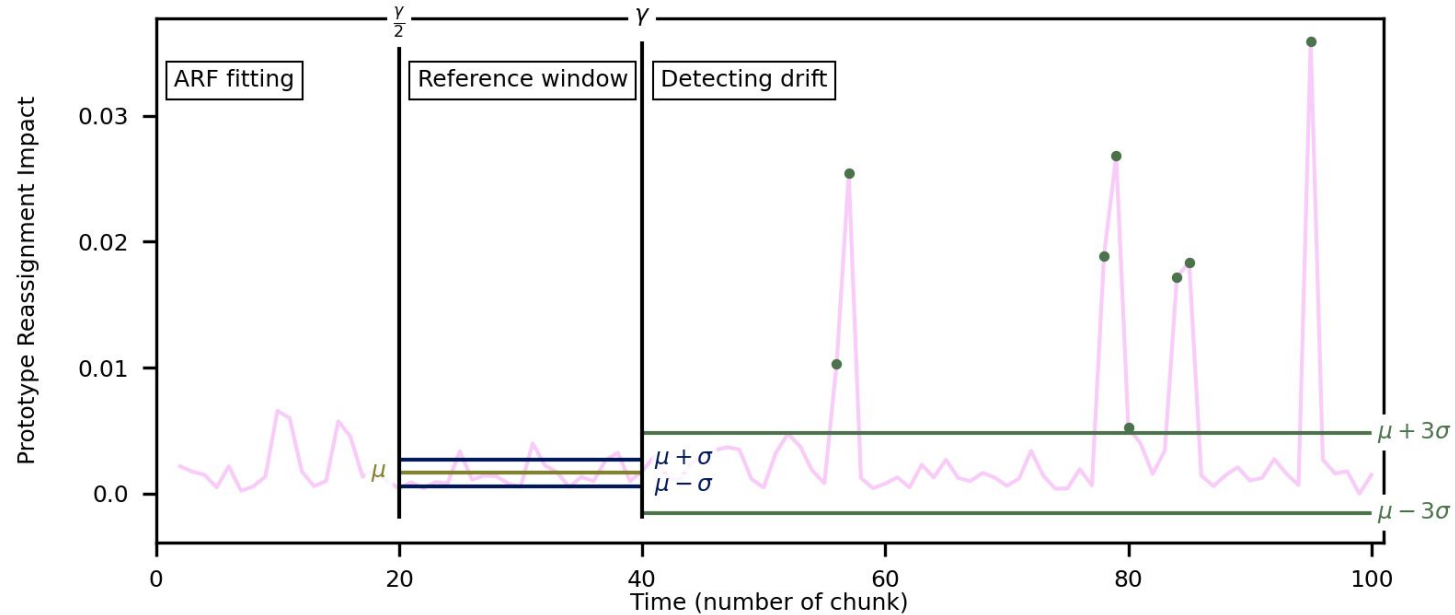
RACE-P: Real-time Analysis of Concept Evolution with Prototypes

Algorithm 1 RACE-P: Real-time Analysis of Concept Evolution with Prototypes

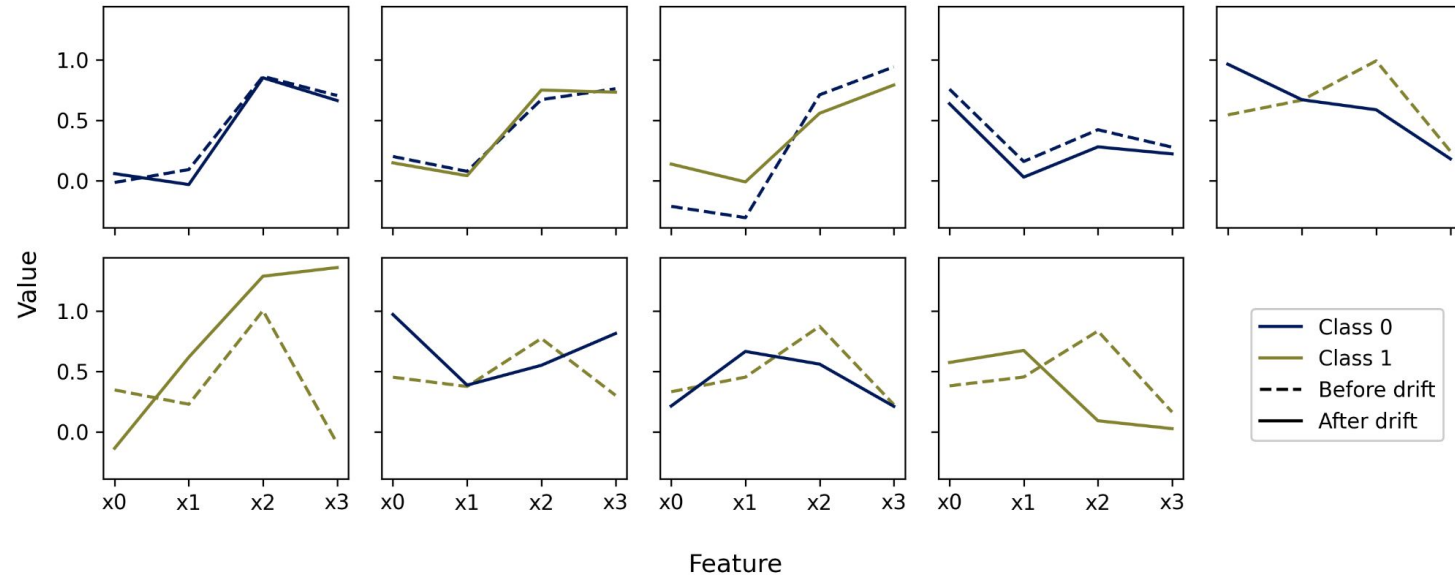
Input: stream of examples \mathcal{S} , grace period γ **Output:** Alarm when a drift occurs

```
1:  $x_{\text{history}}, y_{\text{history}} \leftarrow [\phi, \phi], [\phi, \phi]$ 
2:  $e, p \leftarrow \phi, \phi$ 
3:  $\Delta_{\text{sample}} \leftarrow []$ 
4: ARF  $\leftarrow$  initialize Adaptive Random Forest
5: for  $i \in \{0 \dots |\mathcal{S}| - 1\}$  do
6:   ARF.learn_chunk( $x_i, y_i$ )
7:    $x_{\text{history}}, y_{\text{history}} \leftarrow [x_{\text{history}}[1], x_i], [y_{\text{history}}[1], y_i]$ 
8:    $e \leftarrow [e[1], \text{A-PETE}(\text{ARF})]$ 
9:    $p \leftarrow [p[1], e[1].\text{find\_prototypes}(x_i, y_i)]$ 
10:  if  $i \geq 1$  then
11:     $\Delta \leftarrow \text{PRI}(p[0], p[1])$ 
12:    if  $\frac{\gamma}{2} \leq i < \gamma$  then ▷ Reference Window Creation
13:       $\Delta_{\text{sample}}.\text{append}(\Delta)$ 
14:    if  $i = \gamma$  then
15:       $\mu_{\Delta}, \sigma_{\Delta} \leftarrow \text{avg}(\Delta_{\text{sample}}), \text{stddev}(\Delta_{\text{sample}})$ 
16:      if  $(i \geq \gamma) \wedge (\Delta > \mu_{\Delta} + 3\sigma_{\Delta})$  then ▷ Drift Detection
17:        alarm("Drift detected!")
```

RACE-P: Real-time Analysis of Concept Evolution with Prototypes



Interpreting prototypes



Taxonomy

Prototype-based explanations

```
graph TD; A[Prototype-based explanations] --> B[ante-hoc]; A --> C[post-hoc]
```

ante-hoc

Integrated during model training

Intrinsically interpretable model

Ensures interpretability by design

Often uses **parts or patches** as prototypes

post-hoc

Generated after model training

Any black-box model

Aims to **explain** an **existing model**

Often uses **entire instances** as prototypes



What to focus on?

Training instances often are described by numerous features, but only a subset contributes meaningfully to model behavior or prototype similarity.

Guiding user attention to important parts

1. Find the prototype nearest to the instance to explain.
2. Compute a feature mask highlighting features important for both the prototype and the explained distance:

$$w_l = \hat{\phi}(h, \mathbf{x}_i^l) \cdot \hat{\phi}(h, \mathbf{p}_j^l)$$

$$m_l = \mathbb{1} \left(w_l > \frac{1}{d} \sum_{k=1}^d w_k \right)$$

	Size	Weight	Sweetness	Crunchiness	Juiciness	Ripeness	Acidity
Instance	-2.77	-1.08	-1.72	1.38	0.19	3.65	0.31
Prototype	-0.97	-0.20	-3.07	0.00	-0.52	3.16	-0.52
Weights	0.18	0.02	0.27	0.00	0.00	0.51	0.00
Mask	1	0	1	0	0	1	0

[7] Karolczak, J. and Stefanowski, J. 2025. This part looks alike this: identifying important parts of explained instances and prototypes. In proceedings of The 3rd World Conference on eXplainable Artificial Intelligence

[10] Lundberg, S. and Lee, S. 2017. A unified approach to interpreting model predictions. In the Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 4768-4777. doi:10.5555/3295222.3295230

Guiding user attention to important parts

1. Find the prototype nearest to the instance to explain.
2. Compute a feature mask highlighting **features important** for both the prototype and the explained distance:

$$w_l = \hat{\phi}(h, \mathbf{x}_i^l) \cdot \hat{\phi}(h, \mathbf{p}_j^l)$$

$$m_l = 1 \left(w_l > \frac{1}{d} \sum_{k=1}^d w_k \right)$$



SHAP

	Size	Weight	Sweetness	Crunchiness	Juiciness	Ripeness	Acidity
Instance	-2.77	-1.08	-1.72	1.38	0.19	3.65	0.31
Prototype	-0.97	-0.20	-3.07	0.00	-0.52	3.16	-0.52
Weights	0.18	0.02	0.27	0.00	0.00	0.51	0.00
Mask	1	0	1	0	0	1	0

[7] Karolczak, J. and Stefanowski, J. 2025. This part looks alike this: identifying important parts of explained instances and prototypes. In proceedings of The 3rd World Conference on eXplainable Artificial Intelligence

[10] Lundberg, S. and Lee, S. 2017. A unified approach to interpreting model predictions. In the Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 4768-4777. doi:10.5555/3295222.3295230

Feature Importance is not the only option

In fact, post-hoc feature importance estimators represents only one perspective. The majority of existing methods rely on alternative strategies.



Taxonomy

Prototype-based explanations



```
graph TD; A[Prototype-based explanations] --> B[ante-hoc]; A --> C[post-hoc]
```

ante-hoc

post-hoc

Integrated during model training

Generated after model training

Intrinsically interpretable model

Any black-box model

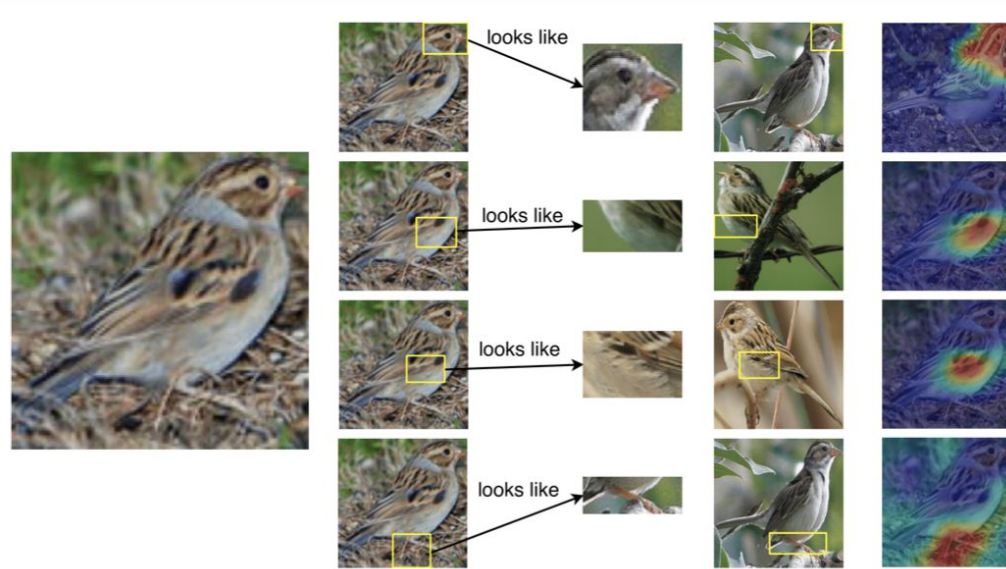
Ensures interpretability by design

Aims to **explain** an **existing model**

Often uses **parts or patches** as prototypes

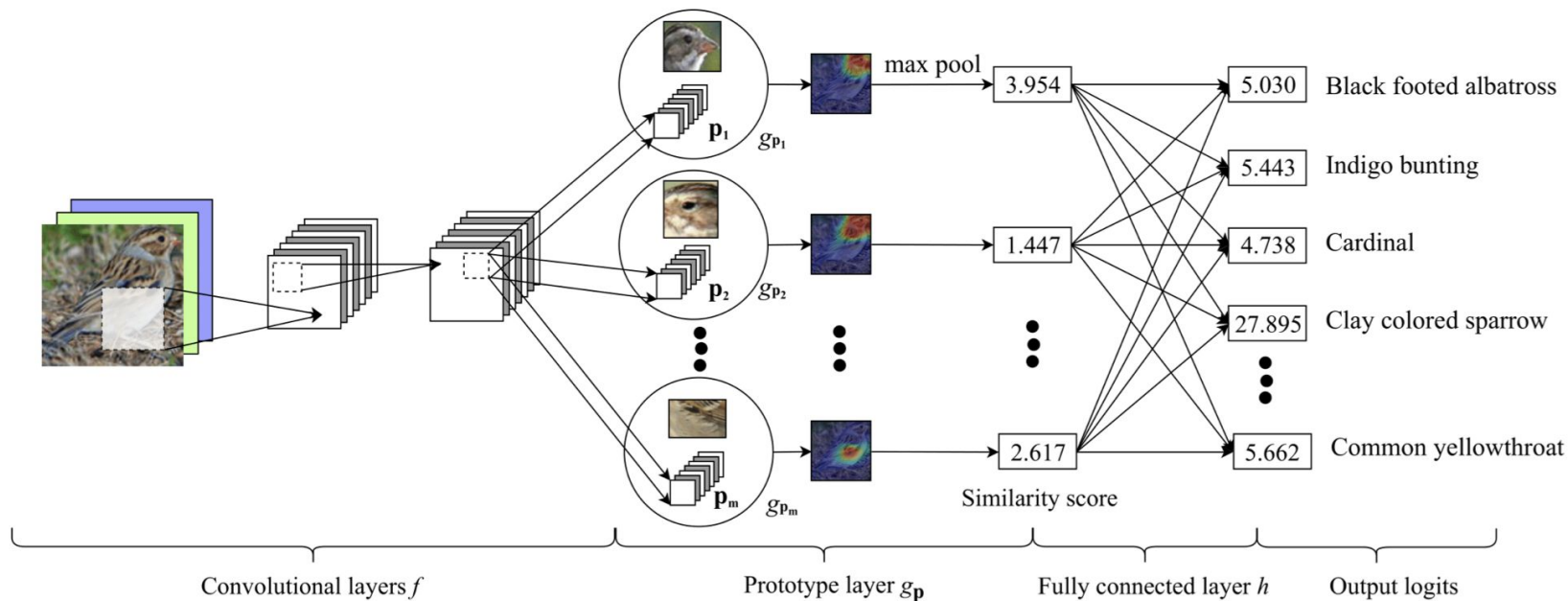
Often uses **entire instances** as prototypes

ProtoNet: Prototypical Part Network



[3] Chen, C., Li, O., Tao, C., Barnett A. J., Su, J. and Rudin, C. 2019. This Looks like That: Deep Learning for Interpretable Image Recognition. In Proceedings of the 33rd International Conference on Neural Information Processing Systems.

ProtoNet: Prototypical Part Network



[3] Chen, C., Li, O., Tao, C., Barnett A. J., Su, J. and Rudin, C. 2019. This Looks like That: Deep Learning for Interpretable Image Recognition. In Proceedings of the 33rd International Conference on Neural Information Processing Systems

Typical prototypical-part network training

1 Initial training

Train the network to learn artificial prototypical parts that capture representative part-based features.

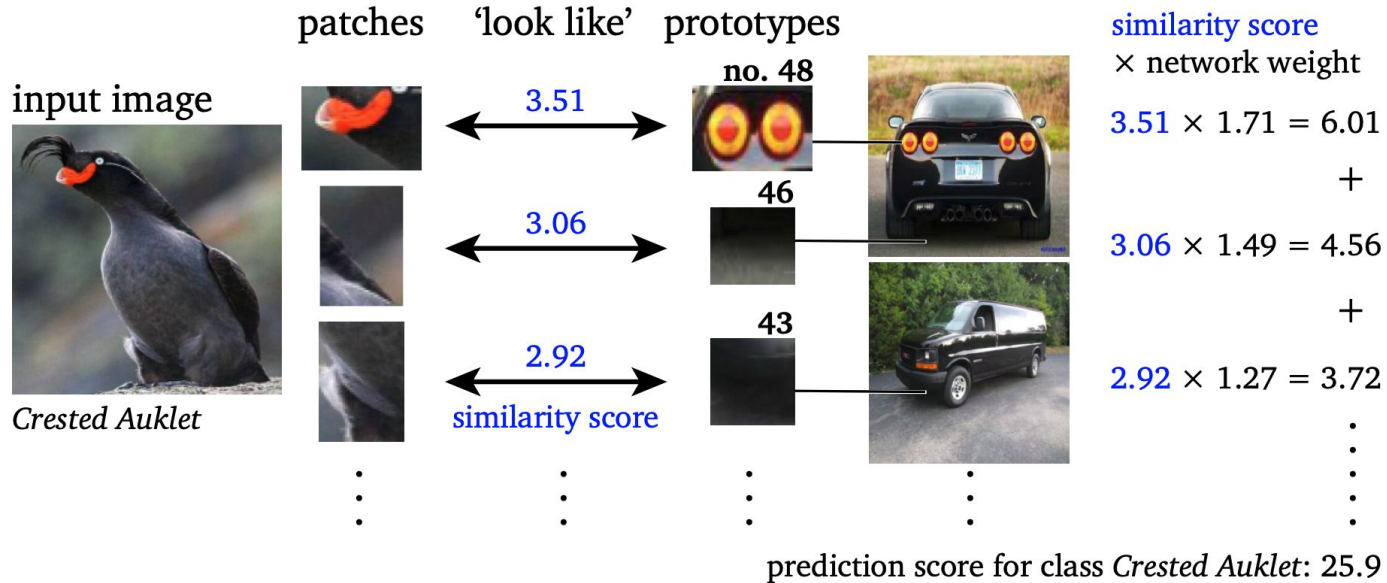
2 Prototype replacement

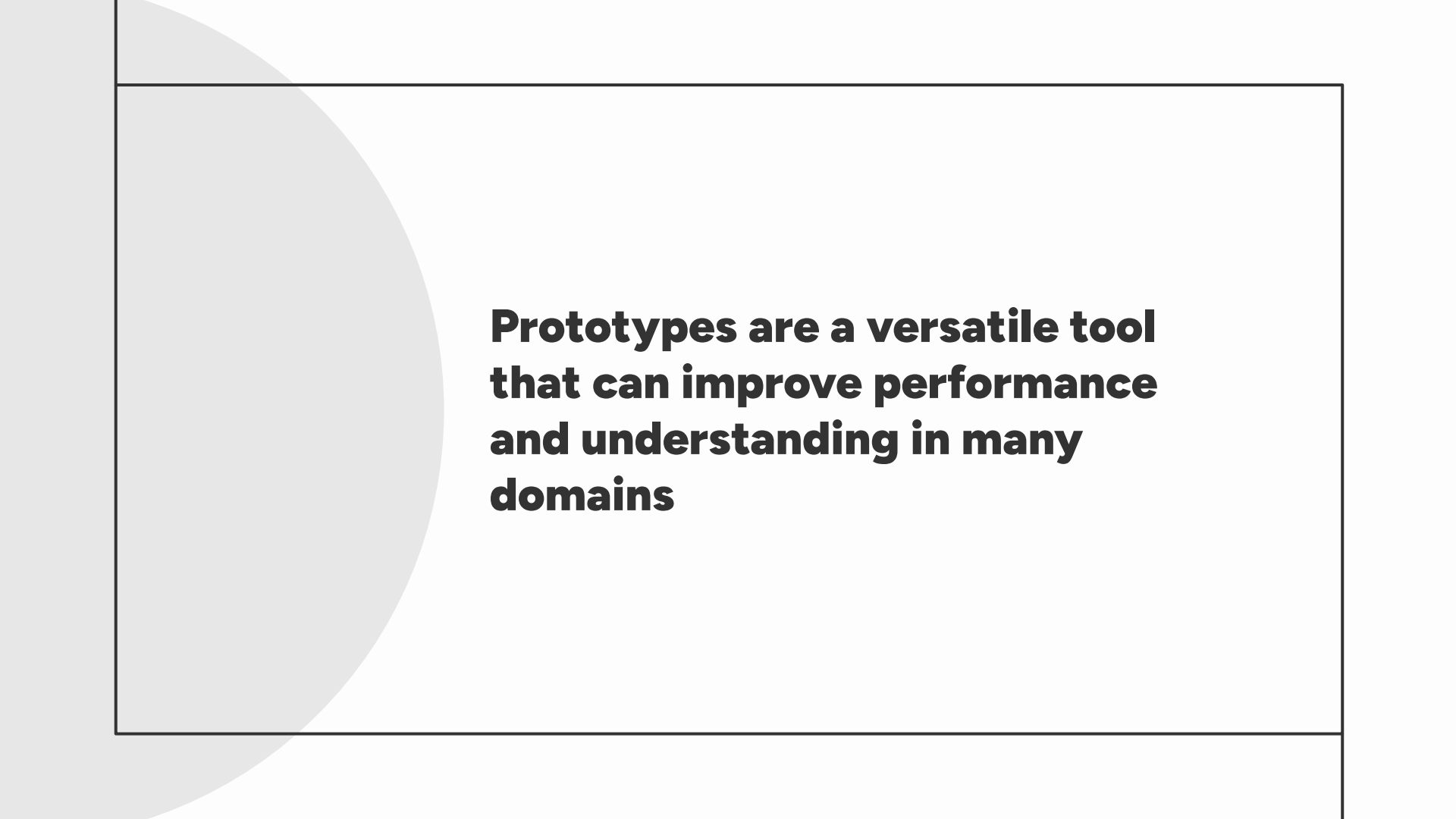
Replace the artificial prototypes with their nearest real parts from the dataset to ground the model in real-world features.

3 Fine-tuning

Fine-tune the classification head using the real prototypical parts
To improve final prediction accuracy

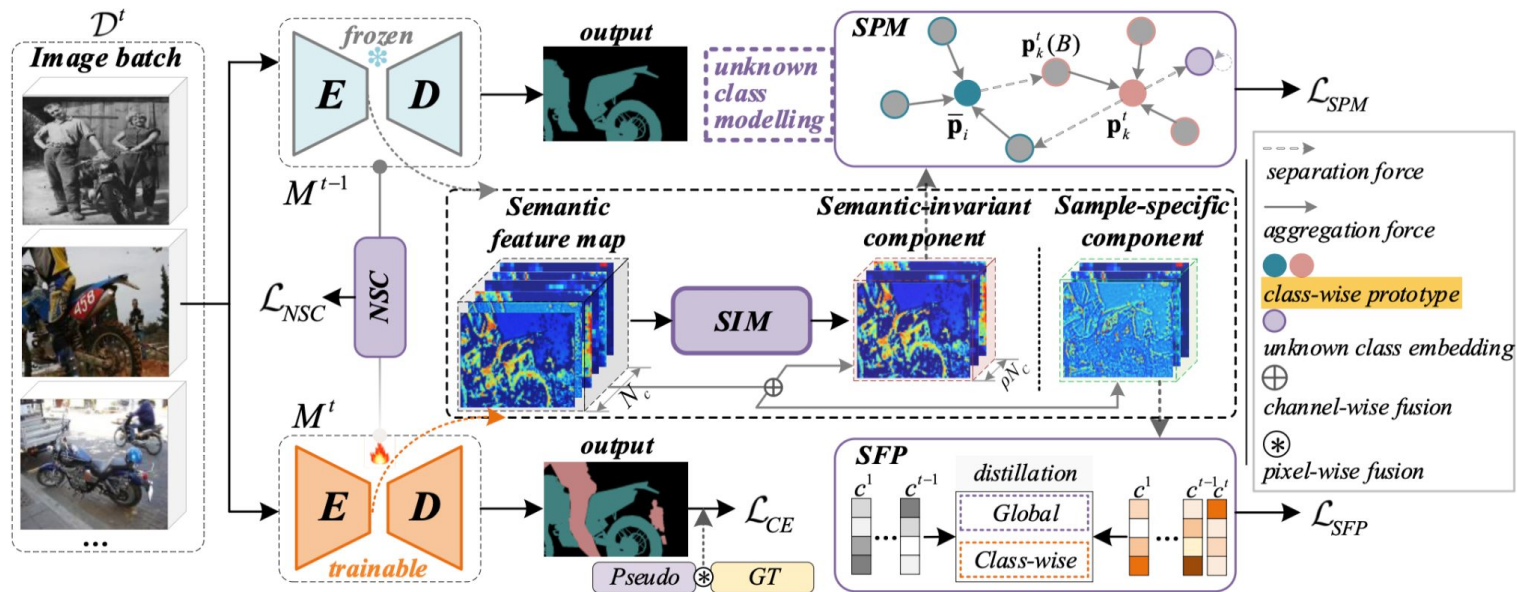
When birds look like cars and other issues with the approach



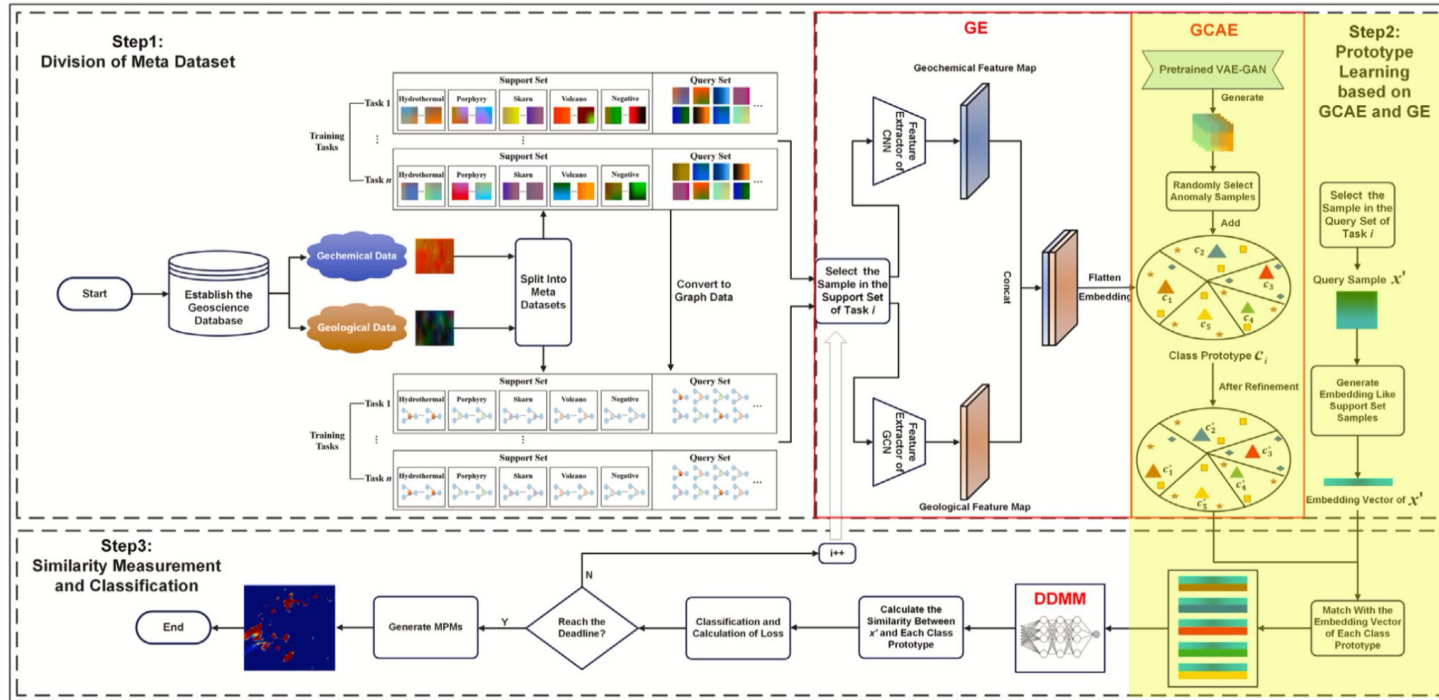


**Prototypes are a versatile tool
that can improve performance
and understanding in many
domains**

Learning at a Glance: Towards Interpretable Data-Limited Continual Semantic Segmentation via Semantic-Invariance Modelling



Mineral prediction based on prototype learning



Key takeaways

01

Explanations should be meaningful, not just numerically sound.

02

Prototypes (can) offer intuitive, concept-level insights.

03

Do not overwhelm users – guide their attention to what really matters.

04

Similarity is not just about raw features – it is about internal model reasoning.

05

Using prototypes may not only support interpretability, but also help with other challenges.



Listening carefully to the entire presentation



Taking a photo of the final takeaway slide

Bibliographical references

- [1] Baniecki, H. and Biecek, P. Birds look like cars: Adversarial analysis of intrinsically interpretable deep learning. 2025. Available at arXiv. doi:10.48550/arXiv.2503.08636
- [2] Breiman, L., 2001. Random forests. *Machine Learning* 45, 5–32. doi:10.1023/A:1010933404324
- [3] Chen, C., Li, O., Tao, C., Barnett A. J., Su, J. and Rudin, C. 2019. This Looks like That: Deep Learning for Interpretable Image Recognition. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 8930-8941. doi:10.5555/3454287.3455088
- [4] Ding, L., Chen, B., Zhu, Y., Dong, H., Zhang, P. 2024. Mineral prediction based on prototype learning. In *Computers & Geosciences*, vol. 184, 105540, doi: 10.1016/j.cageo.2024.105540
- [5] Karolczak, J. and Stefanowski, J. 2024. A-PETE: Adaptive Prototype Explanations of Tree Ensembles. *Progress in Polish Artificial Intelligence Research 5 : Proceedings of the 5th Polish Conference on Artificial Intelligence (PP-RAI'2024)*. p. 2-8. doi:10.17388/WUT.2024.0002.MiNI
- [6] Karolczak, J. and Stefanowski, J. 2025. Explaining Data Changes with Prototypes: A Measure-Driven Approach. Available at SSRN. doi:10.2139/ssrn.5208587
- [7] Karolczak, J. and Stefanowski, J. 2025. This part looks alike this: identifying important parts of explained instances and prototypes. In *proceedings of The 3rd World Conference on eXplainable Artificial Intelligence*
- [8] Kraus, A., van der Aa, H. 2025. Machine learning-based detection of concept drift in business processes. *Process Sci* 2, 5. doi: 10.1007/s44311-025-00012-w
- [9] Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Ser, J. D., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Paez, A., Samek, W., Schneider, J., Speith, T. and Stumpf, S. 2024. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *An International Journal on Information Fusion*, 106(102301), 102301. doi:10.1016/j.inffus.2024.102301
- [10] Lundberg, S. and Lee, S. 2017. A unified approach to interpreting model predictions. In the *Proceedings of the 31st International Conference on Neural Information Processing Systems*. pp. 4768-4777. doi:10.5555/3295222.3295230
- [11] Samek, W., Binder, A., Lapuschkin, S. and Müller, K.-R. 2017. Understanding and Comparing Deep Neural Networks for Age and Gender Classification. 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 1629–1638. doi:10.1109/ICCVW.2017.191
- [12] Speith, T. 2022. How to evaluate explainability – a case for three criteria. *Proceedings of the 30th IEEE International Requirements Engineering Conference Workshops, REW 2022*. pp. 92-97. doi:10.1109/REW56159.2022.00024
- [13] Yuan, B., Zhao, D., and Shi, Z. 2024. Learning at a Glance: Towards Interpretable Data-Limited Continual Semantic Segmentation via Semantic-Invariance Modelling. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 7909-7923, doi: 10.1109/TPAMI.2024.3396809

Thanks!

Do you have any questions?

jacek.karolczak@cs.put.poznan.pl

CREDITS: This presentation template was created by Slidesgo, and includes icons, infographics by Freepik